Data Help Desk

Get Answers to Your Data Management Questions:
Workshops, Demos and More ...

Data Management Skillbuilding Hub

dataoneorg.github.io/Education/
SESSION 2: UPCOMING ACTIVITIES AND ROAD MAP

Matthew B. Jones, Amber Budden, Dave Vieglais, Lauren Walker, Chris Jones, Robyn Thiessen-Bock

The Data Management Skillbuilding Hub contains resources for better data management and is open to community input and update. These resources are adaptable across a range of contexts and intended for use by researchers, teachers, librarians, or anyone who wants to learn better data management practices. Each tile below links to community contributed education materials, such as best practices and lesson plans.

The resources presented on the Data Management Skillbuilding Hub can be updated by users to promote a current, well-maintained, and sustainable educational tool. Learn more about how you can contribute.

Using This Resource
Click individual tiles to learn more and use each resource. You can limit resources by content type and Data Life Cycle stage. Comprehensive information is available in the FAQ.

- Filter by content type: ALL TEACHING MODULE BEST PRACTICE VIDEO
- Filter by stage of the Data Life Cycle: All
Status of Data Management Education

Survey of Ecology Courses

Percent of ecology courses that address and/or teach the data management topics listed

Strasser and Hampton 2012
Challenges in Data Management Training

Survey of Educators

- There is no time to teach data management (n=69) - 51.5%
- It is not my area of expertise (n=53) - 39.6%
- I don't have enough information (n=41) - 30.6%
Training

Leadership in Data Management Education

Lesson 10: Analysis and Workflows

Typical data analyses:
- Data processing: may include selecting a subset of data for analysis, merging multiple data sets, manipulating data for usability, or data transformation

Graphical analysis: makes it easier to see patterns and can aid in the identification of outliers

Statistical analysis: conventional statistics are used to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data

Science is iterative: the process that results in the final product can be complex.

Reproducibility...

...is at the core of the scientific process. If results are not reproducible, they lose credibility.

Good documentation of the data and the analysis are essential!

Formal Workflow

Formal workflow example: Kepler software

Best practices for data analysis

Formally or informally document the workflows used to create results. Include:
- Data provenance
- Analyses and parameters used
- Connections between analyses via inputs and outputs

Document the code you write for analyses.
- Well-documented code is easier to review and share and enables repeated analyses
- Include project level information; script dependencies, inputs, and outputs; parameters; and what happens in individual sections

Construct end-to-end scripts that run the entire process from start to finish without intervention.

40 Webinars

3421 Unique Webinar Attendees

Liz Ferguson

Greg Wilson

Stephanie Hampton

Fernando Pérez
Training

Training Events
<table>
<thead>
<tr>
<th>Quote</th>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Thanking my lucky stars for @DataONEorg education modules this week! Data entry exercise adapted nicely for online participation.”</td>
<td>@DataONEorg</td>
</tr>
<tr>
<td>“Great collection of free education material on data management by @DataONEorg.”</td>
<td></td>
</tr>
<tr>
<td>“Thank you! I love these materials.”</td>
<td>@DataONEorg</td>
</tr>
<tr>
<td>“@DataONEorg website has great education resources on managing data with best practices.”</td>
<td></td>
</tr>
<tr>
<td>“Thanks @DataONEorg for making these education modules available CCO!”</td>
<td></td>
</tr>
<tr>
<td>“Awesome presentations about data management. The Metadata module is great!! I like them. Thanks @DataONEorg”</td>
<td></td>
</tr>
</tbody>
</table>
The Evolution of Resources

1. Objectives
   1. Data management
   2. Efficient data retrieval
   3. Interoperable data

2. Why Manage Data?
   1. It will benefit you and your collaborators.
   2. It will improve the value of your research.
   3. It will ease the process of data retrieval.

Data management has two main components:

- Data cataloging
- Storage and management of the data

Data lifecycle:

1. Capture
2. Organize
3. Store
4. Maintain
5. Retrieved
6. Utilize
7. Disseminate
8. Retain

Data cataloging:

- The identification and description of data assets
- The maintenance of metadata about the data assets

Data management tools:

- DataONE
- NCEAS
- DataONE Education Modules

DataONE Education Modules:

- Data Stewardship and Access
  - Data Quality Control and Assurance
  - Analysis and Workflows

Lesson 10: Analysis and Workflows

Typical data analyses:

- Data processing: selecting a subset of data, merging multiple data sets, manipulating data for usability, or data transformation
- Graphical analysis: making it easier to see patterns and can aid in the identification of outliers
- Statistical analysis: conventional statistics are used to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data

Science is iterative: the process that results in the final product can be complex.

Reproducibility:

- Is at the core of the scientific process. If results are not reproducible, they lose credibility.
- Good documentation of the data and the analysis are essential!

Formal Workflow:

- Analytical pipeline where each step can be implemented in different software systems.
- Parameters and requirements for each step are formally recorded.
- Single access point for multiple analyses across software packages.
- Keeps track of analysis and provenance to better enable reproducibility.
- Workflow can be stored.
- Allows sharing and reuse of individual steps or overall workflow.

Informal Workflow:

- Precise description of the procedures used in a project. Can be formal or informal.
- No special software is needed to create workflow diagrams.
- Workflow diagrams include:
  - Inputs and outputs
  - Transformation rules or analytical processes
  - Decision points
  - Arrows indicating direction of process flow

Informal Workflow Example:

No special software is needed to create workflow diagrams. Informal diagrams:

- Integrated and analyze
- Examine
- Review
- Share

Formal workflow example: Kepler software

Best practices for data analysis:

- Formally or informally document the workflows used to create results.
- Include:
  - Data provenance
  - Analyses and parameters used
  - Connections between analyses via inputs and outputs
- Document the code you write for analyses.
  - Well-documented code is easier to review and share and enables repeated analyses.
  - Include project-level information; script dependencies, inputs, and outputs; parameters; and what happens in individual sections.
- Construct end-to-end scripts that run the entire process from start to finish without intervention.
The Evolution of Resources

Lesson 10: Analysis and Workflows

Typical data analyses may include selecting a subset of data for analysis, merging multiple data sets, manipulating data for usability, or data transformation. Graphical analysis makes it easier to see patterns and can aid in the identification of outliers. Statistical analysis uses conventional statistics to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data.

Science is iterative: the process that leads to the final product can be complex. Reproducibility is at the core of the scientific process. If results are not reproducible, they lose credibility. Good documentation of the data and the analysis are essential!

Formal Workflow

Formal workflow is implemented in different software systems. Parameters and requirements for each step are formally recorded. A single access point for multiple analyses across software packages keeps track of analysis and provenance to better enable reproducibility. A workflow can be stored and allows sharing and reuse of individual steps or overall workflow.

Workflows

Definition: Precise description of the procedures used in a project. Can be formal or informal. Informal workflow does not need special software to create workflow diagrams. Workflow diagrams include:
- Inputs and outputs
- Transformation rules or analytical processes
- Decision points
- Arrows indicating direction of process flow

Informal Workflow Example

Formal workflow example: Kepler software

Best practices for data analysis

Formally or informally document the workflows used to create results. Include:
- Data provenance
- Analyses and parameters used
- Connections between analyses via inputs and outputs

Document the code you write for analyses:
- Well-documented code is easier to review and share and enables repeated analyses
- Include project-level information: script dependencies, inputs, and outputs; parameters; and what happens in individual sections

Construct end-to-end scripts that run the entire process from start to finish without intervention.

Education Materials
The Data Management Skillbuilding Hub contains resources for better data management and is open to community input and update. These resources are adaptable across a range of contexts and intended for use by researchers, teachers, librarians, or anyone who wants to learn better data management practices. Each tile below links to community contributed education materials, such as best practices and lesson plans.

The resources presented on the Data Management Skillbuilding Hub can be updated by users to promote a current, well-maintained, and sustainable educational tool. Learn more about how you can contribute.
When first sharing research data, researchers often raise questions about the value, benefits, and mechanisms for sharing. Many stakeholders and interested parties, such as funding agencies, communities, other researchers, or members of the public may be interested in research, results and related data. This lesson addresses data sharing in the context of the data life cycle, the value of sharing data, concerns about sharing data, and methods and best practices for sharing data.

Cite this lesson:

Hosted by DataONE
In collaboration with the community, DataONE has developed high quality resources for helping educators and librarians with training in data management, including teaching materials, webinars and a database of best-practices to improve methods for data sharing and management.
Best Practice: Describe

Select a Best Practice below to learn more about the "Describe" stage in the Data Life Cycle.

What is the "Describe" stage?

Document data by describing the why, who, what, when, where, and how of the data. Metadata, or data about data, are key to data sharing and reuse, and many tools such as standards and software are available to help describe data.

More information can be found in the Best Practices Primer.

Assign descriptive file names

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number.(click for more)

Tags: access, describe, discover, format

Choose and use standard terminology to enable discovery

Terms and phrases that are used to represent categorical data values or for creating content in metadata records should reflect appropriate and accepted vocabularies in your community or institution. Methods used to identify and select the proper terms.(click for more)

Tags: controlled vocabulary, describe, documentation, metadata, ontologies, preserve, standards

Confirm a match between data and their description in metadata

To assure that metadata correctly describes what is actually in a data file, visual inspection or analysis should be done by someone not otherwise familiar with the data and its format. This will assure that the metadata is sufficient to describe the data.(click for more)

Tags: assure, data consistency, describe, documentation, metadata, quality

Create a data dictionary

A data dictionary provides a detailed description for each element or variable in your dataset and data model. Data dictionaries are used to document important and useful information such as a descriptive name, the data type, allowed values, units, and ...(click for more)

Tags: controlled vocabulary, describe, documentation, metadata, terminology, units
Data Management Skillbuilding Hub

Guidelines for contributors and content editors

This document details our recommended processes to update current content, suggest changes to content, and fork content for your own use, as well as an introduction to how the content is organized and the tools we use to display content.

This repository was developed by the DataONE Community Engagement and Outreach Working Group and continues to be maintained by members of this team. Thank you for your interest in contributing to these educational materials.

Update current content

Want to update a link or method? See a spelling error? Changes can be easily proposed by opening the GitHub Education page and editing content directly. For help, try this brief GitHub tutorial on forking and editing content.

Edit content
1. Create a fork of the lessons or best practices repository into your github account, depending on which content you wish to edit.
2. Modify the files that you want to change (see “Structure” below for tips on making changes).
3. Submit a pull request against the master branch of this repository.
4. Your changes will be reviewed by the repository admins.

Page not rendering?
Check that the title field of the YAML header (the first line of each lesson) is in quotes.

Suggest changes to content
1. Open an issue on this repository.
2. Provide your suggested changes with as much detail and guidance as possible. Be specific.
3. Your suggestions will be reviewed by the repository admins.
4. Changes will be pushed to the repository by the repository admins regularly as needed.

Fork content for your own use
Fork and edit content through GitHub, rather than editing privately, to enable others to use your edited content and to track how these materials are used.
1. Create a fork of the lessons or best practices repository into your github account
2. Modify the files that you want to change (see “Structure” below for tips on making changes)
How to write a Best Practice file

Data Life Cycle stage(s): Life cycle stage here

Filename
Each file is named as the first three words of the title, separated by dashes. This page’s title is “How to write a best practice”, therefore the filename is how-to-write.md.

Header
The header (aka: front matter) for best practices has the following fields:
title: the title of the webpage, as it will appear in the lists
layout: is always bestpractice [NB this will likely be done automatically using the collections features from jekyll]
tags: a list of (short) keywords describing the content of the best practice text
step: a list of one or more steps of the data lifecycle to which this best practice applies
related: a list of related best practices identifiers – an identifier is the first three words of the title, separated by dashes (optional)
updated: the date this best practice was created
author: a list of authors that created the best practice
organization: name of organization that oversaw the creation of the best practice
org_url: website of the organization, organization logo will open this webpage when selected
org_logo: name of the organization’s logo file, this must be a png
categories: this must be listed as “[Best Practice]”, used for sorting and accessing education materials

in ‘Raw’ view, these are the headers you should have

content
Data Management Skillbuilding Hub

Edit or fork content for your own use

Select one of the buttons below to open the appropriate GitHub repository:

- Teaching Module
- Best Practice

Hosted by DataONE

In collaboration with the community, DataONE has developed high-quality resources for helping educators and librarians with training in data management, including teaching materials, webinars and a database of best-practices to improve methods for data sharing and management.

If you have a question or concern, please open an Issue in this repository on GitHub.
dataoneorg.github.io/Education/
Reproducible Research Techniques for Synthesis

A five day immersion into widely adopted R-based tools for open science

DataONE  NCEAS
Curriculum at a Glance

Enable data reuse through better data management
- Metadata - how to write a quality data description
- Data modeling - tidy data for efficient access and storage
- Data publishing, citation and credit

Build reproducible scientific workflows
- Data munging with R tidyverse
- Working collaboratively - git and GitHub
- Writing functions in R
- Building packages for publishing replicable research

Communicate results effectively
- Literate analysis with RMarkdown
- Publishing analytical web pages with GitHub pages
- Data visualization with ggplot and leaflet

Details

Dates:
- November 4-8, 2019
- February, 2020 TBD
- May, 2020 TBD

Cost:
- $2,100
  - Includes: 5 days of instruction, coffee, light snacks and lunch.
  - Does not include: travel, lodging costs, breakfast or dinner.

Location:
- NCEAS, 735 State St. Suite #300
- Santa Barbara, CA